

Scheduling and Control of Queueing Networks

Applications of queueing network models have multiplied in the last generation, including scheduling of large manufacturing systems, control of patient flow in health systems, load balancing in cloud computing, and matching in ride sharing. These problems are too large and complex for exact solution, but their scale allows approximation.

This book is the first comprehensive treatment of fluid scaling, diffusion scaling, and many-server scaling in a single text presented at a level suitable for graduate students. Fluid scaling is used to verify stability, in particular treating max weight policies, and to study optimal control of transient queueing networks. Diffusion scaling is used to control systems in balanced heavy traffic, by solving for optimal scheduling, admission control, and routing in Brownian networks. Many-server scaling is studied in the quality and efficiency driven Halfin–Whitt regime and applied to load balancing in the supermarket model and to bipartite matching in ride-sharing applications.

GIDEON WEISS is Professor Emeritus in the Department of Statistics at the University of Haifa, Israel. He has previously held tenured positions at Tel Aviv University and at Georgia Tech Industrial and Systems Engineering and visiting positions at Berkeley, MIT, Stanford, NYU, and NUS. He is author of some 90 research papers and served on the editorial boards of leading journals on operations research and applied probability. His work includes significant contributions to the fields of time series, stochastic scheduling, bandit problems, fluid analysis of queueing networks, continuous linear programming, and matching problems.

INSTITUTE OF MATHEMATICAL STATISTICS
TEXTBOOKS

Editorial Board

Nancy Reid (University of Toronto)
Arnaud Doucet (University of Oxford)
Xuming He (University of Michigan)
Ramon van Handel (Princeton University)

ISBA Editorial Representative

Peter Müller (University of Texas at Austin)

IMS Textbooks give introductory accounts of topics of current concern suitable for advanced courses at master's level, for doctoral students and for individual study. They are typically shorter than a fully developed textbook, often arising from material created for a topical course. Lengths of 100–290 pages are envisaged. The books typically contain exercises.

In collaboration with the International Society for Bayesian Analysis (ISBA), selected volumes in the IMS Textbooks series carry the “with ISBA” designation at the recommendation of the ISBA editorial representative.

Other Books in the Series (*with ISBA)

1. *Probability on Graphs*, by Geoffrey Grimmett
2. *Stochastic Networks*, by Frank Kelly and Elena Yudovina
3. *Bayesian Filtering and Smoothing*, by Simo Särkkä
4. *The Surprising Mathematics of Longest Increasing Subsequences*, by Dan Romik
5. *Noise Sensitivity of Boolean Functions and Percolation*, by Christophe Garban and Jeffrey E. Steif
6. *Core Statistics*, by Simon N. Wood
7. *Lectures on the Poisson Process*, by Günter Last and Mathew Penrose
8. *Probability on Graphs (Second Edition)*, by Geoffrey Grimmett
9. *Introduction to Malliavin Calculus*, by David Nualart and Eulàlia Nualart
10. *Applied Stochastic Differential Equations*, by Simo Särkkä and Arno Solin
11. **Computational Bayesian Statistics*, by M. Antónia Amaral Turkman, Carlos Daniel Paulino, and Peter Müller
12. *Statistical Modelling by Exponential Families*, by Rolf Sundberg
13. *Two-Dimensional Random Walk: From Path Counting to Random Interlacements*, by Serguei Popov

Scheduling and Control of Queueing Networks

GIDEON WEISS
University of Haifa, Israel



CAMBRIDGE
UNIVERSITY PRESS

Cambridge University Press
978-1-108-41532-3 — Scheduling and Control of Queueing Networks
Gideon Weiss
Frontmatter
[More Information](#)

CAMBRIDGE UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom
One Liberty Plaza, 20th Floor, New York, NY 10006, USA
477 Williamstown Road, Port Melbourne, VIC 3207, Australia
314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre,
New Delhi – 110025, India
103 Penang Road, #05–06/07, Visioncrest Commercial, Singapore 238467

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning, and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781108415323
DOI: 10.1017/9781108233217

© Gideon Weiss 2022

This publication is in copyright. Subject to statutory exception and to the provisions of relevant collective licensing agreements, no reproduction of any part may take place without the written permission of Cambridge University Press.

First published 2022

A catalogue record for this publication is available from the British Library.

Library of Congress Cataloging-in-Publication Data

Names: Weiss, Gideon, author.

Title: Scheduling and control of queueing networks / Gideon Weiss,
The University of Haifa.

Description: First edition. | New York : Cambridge University Press, [2021] |

Series: Institute of mathematical statistics textbooks | Includes
bibliographical references and index.

Identifiers: LCCN 2021029970 | ISBN 9781108415323 (hardback)

Subjects: LCSH: Queueing theory. | System analysis—Mathematics. |

Scheduling—Mathematics. | BISAC: MATHEMATICS / Probability & Statistics / General

Classification: LCC QA274.8 .W43 2021 | DDC 519.8/2—dc23

LC record available at <https://lcn.loc.gov/2021029970>

ISBN 978-1-108-41532-3 Hardback

ISBN 978-1-108-40117-3 Paperback

Additional resources for this publication at www.cambridge.org/9781108415323

Cambridge University Press has no responsibility for the persistence or accuracy of URLs for external or third-party internet websites referred to in this publication and does not guarantee that any content on such websites is, or will remain, accurate or appropriate.

Contents

<i>Notation</i>	xi
<i>Introduction</i>	xiii
Part I The Single Queue	1
1 Queues and Their Simulations, Birth and Death Queues	3
1.1 The Single Queue	3
1.2 Simulation of a Queue	6
1.3 Birth and Death Queues	7
1.4 Historical Notes, Sources and Extensions	13
Exercises	13
2 The M/G/1 Queue	16
2.1 Little's Law	16
2.2 Work Conservation	21
2.3 Some Renewal Theory	23
2.4 Length Biasing	25
2.5 Stationary Point Processes and Palm Measure	27
2.6 PASTA — Poisson Arrivals See Time Averages	28
2.7 M/G/1 Average Waiting Times	31
2.8 Busy Periods	33
2.9 Supplementary Material: Embedded Markov Chains	34
2.10 Sources	39
Exercises	39
3 Scheduling	42
3.1 Batch Scheduling	43
3.2 Value of Information	45
3.3 Stream Scheduling	46
3.4 Optimality of SRPT	47

vi	<i>Contents</i>	
3.5	Mean and Variance for Non-Preemptive, Non-Predictive Scheduling	49
3.6	M/G/1 Priority Scheduling	50
3.7	M/G/1 under Preemptive LCFS and under PS	52
3.8	Smith's Rule, Klimov's Control of M/G/1, and the $c\mu$ Rule.	53
3.9	Bandit Processes and the Gittins Index	53
3.10	Sources	55
	Exercises	56
	Part II Approximations of the Single Queue	59
4	The G/G/1 Queue	61
4.1	Loynes Construction for Stability of G/G/1 Queues	61
4.2	G/G/1 and the Random Walk	64
4.3	Bounds for G/G/1 Waiting Times	67
4.4	Sources	68
	Exercises	69
5	The Basic Probability Functional Limit Theorems	71
5.1	Convergence of Stochastic Processes	71
5.2	Functional Limit Theorems for Random Walks	73
5.3	Functional Limit Theorems for Renewal Processes	79
5.4	Sources	81
	Exercises	81
6	Scaling of G/G/1 and G/G/∞	83
6.1	Primitives and Dynamics of G/G/1	83
6.2	Scaling Time, Space, and Initial Conditions	85
6.3	Skorohod's Reflection Mapping for the Single-Server Queue	86
6.4	Fluid Scaling and Fluid Limits of G/G/1	88
6.5	Diffusion Scale Limits for Underloaded and Overloaded G/G/1	90
6.6	Diffusion Scaling of the G/G/1 Queue in Heavy Traffic	93
6.7	Approximations of the G/G/1 Queue	95
6.8	Another Approach: Strong Approximation	97
6.9	Diffusion Limits for G/G/ ∞ in Heavy Traffic	98
6.10	Sources	100
	Exercises	101
7	Diffusions and Brownian Processes	102
7.1	Diffusion Processes, Brownian Motion, and Reflected Brownian Motion	102
7.2	Diffusions and Birth and Death Processes	105

<i>Contents</i>		vii
7.3	Approximation of the G/G/1 Queue	107
7.4	Two Sided Regulation of Brownian Motion	107
7.5	Optimal Control of a Manufacturing System	109
7.6	Oblique Reflection and Multivariate RBM	111
7.7	Supplement: Calculations for Brownian Motion and Derived Diffusions	112
7.8	Sources	120
	Exercises	120
 Part III Queueing Networks		 123
8	Product-Form Queueing Networks	125
8.1	The Classic Jackson Network	125
8.2	Reversibility and Detailed Balance Equations	127
8.3	Partial Balance and Stationary Distribution of Jackson Networks	129
8.4	Time Reversal and the Arrival Theorem	131
8.5	Sojourn Times in Jackson Networks	133
8.6	Closed Jackson Networks	134
8.7	Kelly Networks	135
8.8	Symmetric Queues and Insensitivity	138
8.9	Multi-Class Kelly-Type Networks	140
8.10	Sources	140
	Exercises	141
 9	 Generalized Jackson Networks	 143
9.1	Primitives and Dynamics	143
9.2	Traffic Equations and Stability	144
9.3	Centering and Skorohod Oblique Reflection	146
9.4	Fluid Limits	147
9.5	Diffusion Limits	149
9.6	Stability of Generalized Jackson Networks	153
9.7	Sources	154
	Exercises	154
 Part IV Fluid Models of Multi-Class Queueing Networks		 157
10	Multi-Class Queueing Networks, Instability, and Markov Rep- resentations	159
10.1	Multi-Class Queueing Networks	159
10.2	Some Unstable MCQN	161

viii	<i>Contents</i>	
10.3	Policy Driven Markovian Description of MCQN	165
10.4	Stability and Ergodicity of Uncountable Markov Processes	168
10.5	Sources	173
	Exercises	174
11	Stability of MCQN via Fluid Limits	175
11.1	System Equations, Fluid Limit Model, Fluid Equations and Fluid Solutions	175
11.2	Stability via Fluid Models	181
11.3	Some Fluid Stability Proofs	185
11.4	Piecewise Linear Lyapunov Functions and Global Stability	191
11.5	Sources	194
	Exercises	194
12	Processing Networks and Maximum Pressure Policies	197
12.1	A More General Processing System	198
12.2	Maximum Pressure Policies	201
12.3	Rate Stability Proof via the Fluid Model	204
12.4	Further Stability Results under Maximum Pressure Policy	206
12.5	Applications	209
12.6	Sources	215
	Exercises	215
13	Processing Networks with Infinite Virtual Queues	219
13.1	Introduction, Motivation, and Dynamics	219
13.2	Some Countable Markovian MCQN-IVQ	221
13.3	Fluid Models of Processing Networks with IVQs	226
13.4	Static Production Planning Problem and Maximum Pressure Policies	228
13.5	An Illustrative Example: The Push-Pull System	230
13.6	Sources	233
	Exercises	234
14	Optimal Control of Transient Networks	236
14.1	The Semiconductor Wafer Fabrication Industry	237
14.2	The Finite Horizon Problem Formulation	238
14.3	Formulation of the Fluid Optimization Problem	240
14.4	Brief Summary of Properties of SCLP and its Solution	242
14.5	Examination of the Optimal Fluid Solution	245
14.6	Modeling Deviations from the Fluid as Networks with IVQs	247
14.7	Asymptotic Optimality of the Two-Phase Procedure	250
14.8	An Illustrative Example	252
14.9	Implementation and Model Predictive Control	253

Contents

ix

14.10	Sources	255
	Exercises	255
Part V Diffusion Scaled Balanced Heavy Traffic		257
15	Join the Shortest Queue in Parallel Servers	259
15.1	Exact Analysis of Markovian Join the Shortest Queue	259
15.2	Variability and Resource Pooling	262
15.3	Diffusion Approximation and State Space Collapse	264
15.4	Threshold Policies for Routing to Parallel Servers	266
15.5	A Note about Diffusion Limits for MCQN	266
15.6	Sources	267
	Exercises	268
16	Control in Balanced Heavy Traffic	270
16.1	MCQN in Balanced Heavy Traffic	270
16.2	Brownian Control Problems	273
16.3	The Criss-Cross Network	276
16.4	Sequencing for a Two-Station Closed Queueing Network	281
16.5	Admission Control and Sequencing for a Two-Station MCQN	287
16.6	Admission Control and Sequencing in Multi-Station MCQN	295
16.7	Asymptotic Optimality of MCQN Controls	300
16.8	Sources	301
	Exercises	302
17	MCQN with Discretionary Routing	305
17.1	The General Balanced Heavy Traffic Control Problem	305
17.2	A Simple Network with Routing and Sequencing	306
17.3	The Network of Laws and Louth	309
17.4	Further Examples of Pathwise Minimization	312
17.5	Routing and Sequencing with General Cuts	315
17.6	Sources	318
	Exercises	319
Part VI Many-Server Systems		321
18	Infinite Servers Revisited	323
18.1	Sequential Empirical Processes and the Kiefer Process	323
18.2	Stochastic System Equations for Infinite Servers	325
18.3	Fluid Approximation of Infinite Server Queues	326
18.4	Diffusion Scale Approximation of Infinite Server Queues	327

x	<i>Contents</i>	
18.5	Sources	329
	Exercises	329
19	Asymptotics under Halfin–Whitt Regime	330
19.1	Three Heavy Traffic Limits	330
19.2	M/M/s in Halfin–Whitt Regime	331
19.3	Fluid and Diffusion Limits for G/G/s under Halfin–Whitt Regime	336
19.4	sources	345
	Exercises	345
20	Many Servers with Abandonment	346
20.1	Fluid Approximation of G/G/n+G	346
20.2	The M/M/n+M System under Halfin–Whitt Regime	350
20.3	The M/M/n+G System	352
20.4	Sources	358
	Exercises	358
21	Load Balancing in the Supermarket Model	360
21.1	Join Shortest of d Policy	361
21.2	Join the Shortest Queue under Halfin–Whitt Regime	373
21.3	Approaching JSQ: Shortest of $d(n)$ and Join Idle Queue	381
21.4	Sources	382
	Exercises	383
22	Parallel Servers with Skill-Based Routing	385
22.1	Parallel Skill-Based Service under FCFS	386
22.2	Infinite Bipartite Matching under FCFS	390
22.3	A FCFS Ride-Sharing Model	401
22.4	A Design Heuristic for General Parallel Skill-Based Service	404
22.5	Queue and Idleness Ratio Routing	406
22.6	Sources	409
	Exercises	409
	<i>References</i>	413
	<i>Index</i>	427

Notation

\mathbb{P}	probability	F	distribution of interarrival times
\mathbb{E}	expectation	G	distribution of service times
$\mathbf{1}$	characteristic function	H	distribution of patience times
\mathbb{P}_x	probability from initial state x	V_ℓ	waiting time of customer ℓ
\mathbb{E}_x	expectation from initial state x	\bar{V}	mean waiting time
\mathbb{N}	natural numbers, $0, 1, \dots$	W_ℓ	sojourn time of customer ℓ
\mathbb{Z}	integers	\bar{W}	mean sojourn time
\mathbb{R}	the real line	$\bar{\mathcal{W}}$	mean workload
\mathbb{C}	space of continuous functions	ρ	offered load or traffic intensity
\mathbb{D}	space of functions right continuous with left limits	α_i	exogenous arrival rates
$\mathbf{1}$	vector of 1's	λ_i	total arrival rates
π	stationary probabilities	μ_i	service rates
$\mathcal{A}(t)$	arrival process	$p_{i,j}$	routing probabilities
$\mathcal{D}(t)$	departure process	ν_k	nominal allocation
$Q(t)$	queue length	c_a	coefficient of variation of interarrival times
$\mathcal{W}(t)$	workload process	c_s	coefficient of variation of service time
$\mathcal{S}(t)$	service process	C	constituency matrix
$\mathcal{T}(t)$	busy time	C_i	constituency of server i
$\mathcal{I}(t)$	idle time	A	resource consumption matrix
$\mathcal{J}(t)$	free time	R	input-output matrix
a_ℓ	arrival time of customer ℓ	R^{-1}	work requirement matrix
T_ℓ	interarrival time, $T_\ell = a_\ell - a_{\ell-1}$ (Chapters 1–4)	B_κ	a compact neighborhood of the origin
u_ℓ	interarrival time, $u_\ell = a_\ell - a_{\ell-1}$ (Chapters 5–20)	\mathcal{C}	set or subset of customer types (Chapter 22)
X_ℓ	service requirement of customer ℓ (Chapters 1–4)	\mathcal{S}	set or subset of server types (Chapter 22)
v_ℓ	service requirement of customer ℓ (Chapters 5–20)		

\mathcal{U}	set or subset of customer types unique to some servers (Chapter 22)	HOL	head of the line policy
\mathcal{G}	compatibility graph (Chapter 22)	FCFS	first come first served
$\mathcal{P}(J)$	permutations of the set J	LCFS	last come first served, pre-emptive
\emptyset	the empty set	PS	processor sharing
r.h.s.	right-hand side	FBFS	first buffer first served, in re-entrant line
i.i.d.	independent identically distributed	LBFS	last buffer first served, in re-entrant line
u.o.c.	uniformly on compacts	SPT	shortest processing time
a.s.	almost surely	SEPT	shortest expected processing time
c.o.v.	coefficient of variation	SRPT	shortest remaining processing time
pdf	probability density function	IVQ	infinite virtual queue
cdf	cumulative distribution function	BCP	Brownian control problem
BM	Brownian motion	ED	efficiency driven service
RBM	reflected Brownian motion	QD	quality driven service
RCLL	right continuous with left limits	QED	quality and efficiency driven service
MCQN	multi-class queueing networks	CRP	complete resource pooling
PASTA	Poisson arrivals see time averages	PSBS	parallel skilled based service
BP	busy period	ALIS	assign longest idle server
EFSBP	exceptional first service busy period	SD	server dependent service rates
		QIR	queue and idleness ratio policy

Introduction

Queueing networks are all pervasive; they occur in service, manufacturing, communication, computing, internet and transportation. Much of queueing theory is aimed at performance evaluation of stochastic systems. Extending the methods of deterministic optimization to stochastic models so as to achieve both performance evaluation and control is an important and notoriously hard area of research. In this book our aim is to familiarize the reader with recent techniques for scheduling and control of queueing networks, with emphasis on both evaluation and optimization.

Queueing networks of interest are discrete, stochastic dynamical systems, often of very large size, and exact analysis is usually out of the question. Furthermore, the data necessary for exact analysis is rarely available. Thus, to obtain useful results we are led to use approximations. In this book, our emphasis is on approximations obtained from scaled versions of the systems, and analyzing the limiting behavior when the scale tends to infinity. We will be studying three types of scaling: fluid, diffusion, and many server.

Fluid scaling (Part IV): We count time in units of n and space, expressed by number of items, in units of n . This will be a reasonable model to ask what happens to a system with n items in a time span in which n items are processed. Under fluid scaling, a discrete stochastic system may converge to a deterministic continuous process, its fluid model. Fluid models are used in two ways: first, to answer the question of stability – is the system capable of recovering from extreme situations, in which case it may converge to a stationary behavior. Second, perhaps more exciting, we can use fluid scaling to obtain asymptotically optimal control of transient systems over finite time horizons.

Diffusion scaling (Part V): Space is scaled by units of n , and time by units of n^2 . On this scale a stable system may reach stationary behavior, and reveal the congested elements of the network in balanced heavy traffic. Approximation of these by stochastic diffusion processes can be used to evaluate performance measures. Furthermore, on the diffusion scale we

may formulate and solve Brownian control problems and derive efficient policies.

Many-server scaling (Part VI): Increasingly, recent applications involve systems with many servers and a large volume of traffic. For such systems time is not scaled, but the number of servers and the arrival rates are scaled. These models preserve not just first moment parameters of the fluid scaling and second moment parameters of the diffusion scaling, but the full service time distribution of individual items moving through the system. Many-server models are used to answer staffing-level questions, and to achieve quality of service goals.

The first three parts of the book cover more conventional material, as well as introducing some of the techniques used later. *Part I* covers birth and death queues and the M/G/1 queue, and a chapter on scheduling. *Part II* deals with approximations to G/G/1, introducing fluid and diffusion scaling and many-server G/G/∞. Two chapters, Chapter 5 and Chapter 7 survey some of the essential probability theory background, at a semi-precise level. *Part III* of the book introduces Jackson networks and related queueing networks with product-form stationary distributions, and generalized Jackson networks.

The book is aimed at graduate students in the areas of Operations Research, Operations Management, Computer Science, Electrical Engineering, and Mathematics, with some background in applied probability. It can be taught as a two part course, using the first three parts of the book as a basic queueing course, or it can be taught to students already familiar with queueing theory, where the first three parts are skimmed and the emphasis is on the last three parts. I have taught this material three times, at three different schools, as a PhD-level course in a single semester, though it was somewhat tight to include all of the material of the second half of the book. I tried to make each chapter as self-contained as possible, to enable more flexibility in teaching a course, and to be more useful for practitioners.

Each chapter in the book is followed by a list of sources, and by exercises. Some of the exercises lead to substantial extensions of the material, and I provide references for those. A few problems that require further study and much more effort are in addition marked by (*). A solution manual will accompany the book, and be placed on the book website, www.cambridge.org/9781108415323.

I conclude this introduction with six examples of potential applications that the techniques developed in this book are designed for:

Semiconductor wafer fabrication plant: Given the current state of the plant, how to schedule production for the next six weeks. While this ap-

pears at first to be a deterministic job-shop scheduling problem, optimal schedules never work due to unexpected stochastic interference. In Chapter 14 we formulate this as a control problem of a discrete stochastic transient queueing network. We use fluid scaling to obtain and solve a deterministic continuous control problem, and then track the optimal fluid solution using decentralized control.

Input queued crossbar switches: Scheduling the traffic through ultra-high speed communication switches so as to achieve maximum throughput is solved by a maximum pressure policy as described in Chapter 12.

Joint management of operating theaters in a hospital: To control the flow of patients, surgeons, equipment, and theaters on a long-term basis, this can be formulated as a multi-class queueing network, which is operating in stationary balanced heavy traffic. In Chapters 16 and 17 we use diffusion scaling to formulate and solve such problems as stochastic Brownian control problems, and the optimal solution of the Brownian problem is used to determine policies that use admission limits, choice of routes, scheduling priorities, and thresholds.

Control of a call center: This is modeled as a parallel service system, where types of customers are routed to pools of compatible skilled servers. Here design of the compatibility graph, balancing staffing levels, and maintaining acceptable levels of abandonment need to be determined, based on many-server scaling in Chapters 19 and 20.

Cloud computing and web searching: Balancing the utilization of the servers and controlling the lengths of queues at many servers needs to be achieved with a minimum amount of communication. Asymptotic optimality here is achieved by routing tasks to the shorter of several randomly chosen servers. This is modeled by the so-called supermarket problem, as studied in Chapter 21.

Ride sharing: Drivers as well as passengers become available in a random arrival stream, and have limited patience waiting for a match to determine a confirmed trip. Matching available drivers to passengers according to their compatibility, and dispatching on first come first served is analyzed and used to design regimes of operation in Chapter 22.

In writing this book I benefitted from the help of many colleagues, foremost I wish to thank to Ivo Adan, Onno Boxma, Asaf Cohen, Liron Ravner, Shuangchi He, Rhonda Righter, Dick Serfozo, and Hanqin Zhang for their many useful comments and suggestions, and to my students who kept me in check.

Cambridge University Press
978-1-108-41532-3 — Scheduling and Control of Queueing Networks
Gideon Weiss
Frontmatter
[More Information](#)
